Experiments with Using Semantical Categories in Parsing Systems

W. R. Hogenhout Yuji Matsumoto
Nara Institute of Science and Technology
{marc-h,matsu}@is.aist-nara.ac.jp

One much used method in syntactical analysis is statistical training of a hand-written grammar. It is a well-known problem of statistical selection that it ignores lexical and semantical preferences in disambiguation. We have developed a technique for using semantical preferences in statistical training and present experimental results. We compare the results obtained by using different sources of semantical information.

1 Background

Recent developments in the field of broad coverage parsing show two key developments. First, the traditional stochastical grammars where the probability of a production depends only on the left hand side nonterminal are now considered to be too simple, because they depend only on very local information. Second, instead of focussing on the algorithm, it is becoming clear that the selection of the right information used for learning is at least as important. This has been argued in general [5], in relation to context free grammars [2], and in relation to parsing methods that do not require a grammar [3, 10].

It is very easy to see the fundamental problem of stochastical grammars. In a simple sentence such as *"He asked about climbing professionally."* it is impossible to decide what was professional by just looking at the structure (compare *"He asked about paying impatiently."*) Any method that performs well needs to take words (or particular properties of words) into account.

We have developed a method to train and use stochastical grammars with richer stochastical models. This allows the words in the sentence and their properties to influence the parsing results. The method was already described in, for example, [7, 6]. In this paper we generalize on this method, and give experimental results.

2 The Algorithm

The main equations used for the algorithm are given below. This method has been described in previous work [7, 6], but the algorithm we describe here is more general than what was described in previous work. It is based on the Inside Outside Algorithm [1, 9] with chart-parsing but uses extra information in the edges.

We number the rules in the grammar and we write rule number x as R_x. When the left hand side nonterminal of rule x is p we write R_x^p. We also distinguish the edges by numbering them and write edge number k as e_k. When edge e_k was produced with rule x we write e_k^x.

We assume every edge is given some additional information. This may be any information such as the length of the phrase it covers, the number of children, a property of the headword it covers, or anything else that would contribute to disambiguation. If edge e_k was produced with rule x and received additional information A we write $e_k^{x,A}$.

When the edges $e_{c1}...e_{cn}$ can be used to produce edge e_k^A with rule x we write $\{e_{c1}...e_{cn}\} \rightarrow e_k^{x,A}$.

We calculate the inside and outside probabilities of edges. We write this as $I(e)$ and $O(e)$ respectively. We will also be using the probability of a pattern associated with an edge $P(R_x, A)$.

The initial estimation of counts is done with

$$C^w_{\text{initial}}(R_x, A) = \frac{\text{number of times } e^{x,A} \text{ is used in parses of } w}{\text{number of alternative parses of } w}$$

which allows estimating of initial probabilities using the equation

$$P_{\text{initial}}(R_x^p, A) = \frac{\sum_{w \in \text{corpus}} C^w_{\text{initial}}(R_x^p, A)}{\sum_{y,B,w \in \text{corpus}} C^w_{\text{initial}}(R_y^p, B)} .$$

Using these initial estimates the reestimation process is carried out with the equations

$$I(e_k^{x,A}) = \sum_{\{e_{c1},e_{c2},..,e_{cn}\} \rightarrow e_k^{x,A}} P(R_x, A) \prod_{i=1}^{n} I(e_{ci})$$

and

$$O(e_k) = \sum_{\{e_{c1}...e_{cn},e_k\} \rightarrow e_q^{x,A}} O(e_q^{x,A}) P(R_x, A) \prod_{i=1}^{n} I(e_{ci})$$

where we take the probability of the whole sentence to be

$$P(w) = \sum_{e\text{: nonterminal is S}} I(e) \ .$$

Using the Inside and Outside probabilities the reestimation is done with

$$C^w(R_x, A) = \frac{1}{P(w)} \sum_{e_k^{x,A}} O(e_k) I(e_k)$$

and we calculate new probabilities with

$$P_{\text{new}}(R_x^p, A) = \frac{\sum_{w \in \text{corpus}} C^w(R_x^p, A)}{\sum_{y,B,w \in \text{corpus}} \frac{1}{P(w)} \sum_{p:e_k^{y,B}} O(e_k) I(e_k)}.$$

Here $p : e_k^{y,B}$ means we sum over all edges that were constructed with a rule that has p as left hand side nonterminal.

3 Smoothing

Given a corpus it is desirable to estimate a stochastical model as well as possible. Since statistical models are easily overtrained on a corpus smoothing is an important step in statistical modeling.

The kind of smoothing possible here depends on the sort of information that is added to edges. If we assume A consists of a number of components (that is not necessarily true), say A_1, A_2 and A_3, we can interpolate between these values in the following way[1]

$$\hat{P}(R_x, (A_1, A_2, A_3)) = \lambda_1 \hat{P}(R_x, (A_1, A_2, A_3)) +$$

$$\lambda_2 \sum_i \hat{P}(R_x, (A_1, A_2, A_i)) + \lambda_3 \sum_i \hat{P}(R_x, (A_1, A_i, A_3)) +$$

$$\lambda_4 \sum_i \hat{P}(R_x, (A_i, A_2, A_3)) \ .$$

Naturally one could also smooth with, for example, (A_1, A_i, A_j). In stead of taking constant smoothing values, a better method is making a function $\lambda_i(A_1, A_2, A_3)$ and finding locally optimal values with the EM algorithm such that $\sum_i \lambda_i(A_1, A_2, A_3) = 1$.

In the presentation of our experiments we give concrete examples of components of additional information.

4 Experiments

Some experimental results were published in [6, 7], but those were based on the 1994 beta version of the EDR Japanese corpus. All experiments reported on in this paper have been carried out with the final 1995 version of the EDR Japanese corpus [8]. As a consequence the results are not comparable, because the tree structures of the EDR corpus changed and our parsing system was changed with it.

We used the SAX parser [11] and the grammar developed by Dr. Takeshi Fuchi [4]. The grammar was trained with 8635 sentences of up to 25 Japanese characters. The test data consisted of 1000 sentences which were not used for training. Both the training data and the test data consisted of sentences for which the grammar generated at least one parse that did not have crossing errors against the EDR solution, which was only about one out of every four sentences.

The results of our experiments are in table 1. The model indicates the type of added information. We will explain the various models in this section. PCFG is the same grammar trained without additional information (a regular stochastic grammar). All models were trained with the training data and then used to select one parse from the parses that our context free grammar produces for every test sentence. The brackets that indicate a word in the sentence were included in the test.

The abbreviations in the table should be read as follows. The EM column shows the number of bracket pairs that exactly matched a bracket pair in the corpus. CE is the number of bracket pairs that crossed some bracket pair in the corpus. Spur are the brackets that were not in the corpus, but also did not cross a corpus bracket. 0 CE indicates the number of sentences with 0 crossing errors, 1 CE those with 1, 2 CE those with 2, and 3+ CE those with 3 or more. Br. Acc. is bracket accuracy (EM as a percentage of EM+CE+Spur). Br. Recall is bracket recall (EM as a percentage of the 23000 bracket pairs in the treebank).[2]

Using a 95% confidence interval, a difference in bracket accuracy is significant when it is 0.5% or higher for our results. A difference in bracket recall is significant when it is 0.35% or more.

4.1 Experiment 1

In the first experiment we extended every edge using semantical information indicated by categories in the Bunruigoihyou thesaurus [12]. This thesaurus defines a complete ontology with words at the leafs. A few general classes in the ontology were selected by hand. Every word was looked up in the thesaurus, and the general class covering the leaf of the word was used as additional information.

[1] We use a tilde to indicate smoothed values and a hat to indicate estimated values

[2] The treebank contained exactly 23000 brackets for the test set. This is coincidence

Table 1: Results from Experiments.

Model	EM	CE	Spur	0 CE	1 CE	2 CE	3+ CE	Br. Acc.	Br. Recall
PCFG	21801	1065	2192	593	57	194	156	87.0 %	94.8 %
BGH	21933	927	2200	646	49	167	138	87.5 %	95.4 %
Length	22223	644	2197	743	30	135	92	88.7 %	96.6 %
Edges	22173	713	2180	726	34	134	106	88.4 %	96.4 %

In cases were there is semantical ambiguity a word corresponds to more than one leaf. In such cases the word was replaced with the general category that covers the most of these leafs. (This is a very simple strategy, but gives a sensible choice in many of the cases.)

The information added to every edge consisted of two components: the semantical head of the phrase covered by the left child of the edge, and the semantical head of the phrase covered by the right child of the edge. Edges with only one child received additional information consisting of only one component.

When thinking of a production rule this represents the chance that a nonterminal is rewritten with the rule into other nonterminals with the respective semantical heads of the phrases they cover.

The categories we used were as follows. Note that the leftmost digit indicates the most rough division, the rightmost digit indicates the finest division in Bunruigoihyou.

Categories	Example Meanings
110-116	existence, movement, time
117-119	space, place, shape
12, 130-132	spiritual, language, creativity
133-138	culture, politics, economy
14	products and tools
15	natural phenomena
2	action
3	comparison
4	others

The number of categories we use is very low (9 categories and 2 special categories of unclassified words), but notice that a specific rule often receives only one part of speech as a headword, which effectively creates a separate distributions for different parts of speech. The problem with increasing the number of categories is that every further division of the categories reduces the data available for training per category. It is therefore not possible to split categories only by their semantical properties, the incidence in the corpus has to be considered. We tried finer categories than this, but we did not have good results.

Since we had two components for two-headed nonterminals we used smoothing to correct the parameter estimation. When there are two semantical heads we can interpolate as was described earlier with

$$\begin{aligned}\tilde{P}(R_x,(A_1,A_2)) = \\ \lambda_1\hat{P}(R_x,(A_1,A_2)) + \\ \lambda_2\sum_i \hat{P}(R_x,(A_1,A_i)) + \\ \lambda_3\sum_i \hat{P}(R_x,(A_i,A_2))\end{aligned}$$

where A_1 is the left head and A_2 is the right head. In this experiment the parameters best results were obtained with $\lambda_1 = 1$ or some high value. (There was however not much variation).

Table 1 gives the results obtained with the algorithm using Bunruigoihyou classes (BGH), and the stochastical grammar. There is a clear improvement over the stochastic grammar.

4.2 Experiment 2

The second experiment was much more simple than the first. The additional information consisted of the length of the phrase covered by the edge. It had again two components for the nodes with two childs, namely the length of the phrase covered by the left child and the length of the phrase covered by the right child.

Once again we had two components, so we smoothed the data with the equation

$$\begin{aligned}\tilde{P}(R_x,(\text{left length},\text{right length})) = \\ \lambda_1\hat{P}(R_x,\text{left length},\text{right length}) + \\ \lambda_2\sum_i \hat{P}(R_x,(\text{left length},i)) + \\ \lambda_3\sum_i \hat{P}(R_x,(i,\text{right length}))\ .\end{aligned}$$

The model performed best with λ_1 very small and λ_2 and λ_3 both close to 0.5. The results were better than the first experiment, see table 1.

We used the length of the phrase in (Japanese) *characters*, not in words. We have also tried the same process using the number of words rather than the number of characters, but this was less successful than the length in characters.

4.3 Experiment 3

In the third experiment we used the number of subedges covered by the edge. In this case the typical high nonterminals receive higher numbers, and the typically low nonterminals (such as small noun phrases) usually receive low numbers. We divided this between the number of left-side subedges and right-side subedges, and we used the same kind of smoothing as in the previous experiments.

This enables the model to see differences between long and short phrases and sentences. For a short sentence the probability distribution will change because the edge has fewer subedges. This information is different from the second experiment in that the number of edges is only indirectly related to the length of the covered phrase. It did not give the best results, but they were close to those in the second experiment. However, the difference with the second experiment is smaller than our confidence interval (see earlier), so we cannot safely conclude there is a significant difference between experiment 2 and 3.

5 Conclusion

When we started the experiments we expected the semantical headwords to be the most successful. We found however that phrase length in characters had more interesting statistical properties than the semantical categories. The number of subedges also gave more information than the semantical headwords. The differences between the experiment based on phrase length and the experiment based on the number of nodes are too small to draw conclusions.

It may seem logical to try the first experiment with the semantical markers in the EDR corpus in stead of using a thesaurus. But both in this and in previous experiments we were not able to reduce the semantical markers to a small number of broad categories and the results of experiments were disappointing.

This experiment shows the importance of trying various features about phrases and discover which are the most useful to syntactical disambiguation. Since it is impossible to know this through introspection, being able to measure the amount of information conveyed in a certain feature is a key technique. The algorithm we present makes it possible to measure this.

References

[1] J. K. Baker. Trainable grammars for speech recognition. *Speech Communication Papers for the 97th Meeting of the Acoustical Society of America*, pp. 547–550, 1979.

[2] E. Black, F. Jelinek, J. Lafferty, and D. M. Magerman. Towards history-based grammars: Using richer models for probabilistic parsing. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pp. 31–37, 1993.

[3] R. Bod. Using an annotated corpus as a stochastic grammar. In *Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics*, pp. 37–44, 1993.

[4] Takeshi Fuchi. *New Means to Analyze Japanese Morphemes and Dependency Structure and Formalization of Rules to Derive Implied Meanings.* PhD thesis, Tokyo University, 1994. in Japanese.

[5] D. Hindle and M. Rooth. Structural ambiguity and lexical relations. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp. 229–236, 1991.

[6] W. R. Hogenhout. Enrichment of models for stochastic grammars: Semantical categories. Master's thesis, Nara Institute of Science and Technology, 1995. NAIST-IS-MT9451207.

[7] W. R. Hogenhout and Y. Matsumoto. Training stochastical grammars on semantical categories. In *Proceedings of the IJCAI Workshop on New Approaches to Learning for Natural Language Processing*, pp. 65 – 70, Aug. 1995.

[8] Japan Electronic Dictionary Research Institute, Ltd. *EDR Electronic Dictionary Technical Guide*, 1995.

[9] K. Lari and S. J. Young. Applications of stochastic context-free grammars using the Inside-Outside Algorithm. *Computer Speech and Language*, Vol. 5, pp. 237–257, 1991.

[10] D. M. Magerman. Statistical decision-tree models for parsing. In *Proceedings of the 33d Annual Meeting of the Association for Computational Linguistics*, 1995.

[11] Yuji Matsumoto and R. Sugimura. A parsing system based on logical programming. In *Proceedings of the 10th International Joint Conference on Artificial Intelligence*, pp. 671–674, 1987.

[12] National Language Research Institute Publications. *Word List by Semantic Principles (Bunruigoihyou).* Shuei Shuppan, 1964. in Japanese.